An Introduction to Statistical Extreme Value Theory

Uli Schneider

Geophysical Statistics Project, NCAR

January 26, 2004

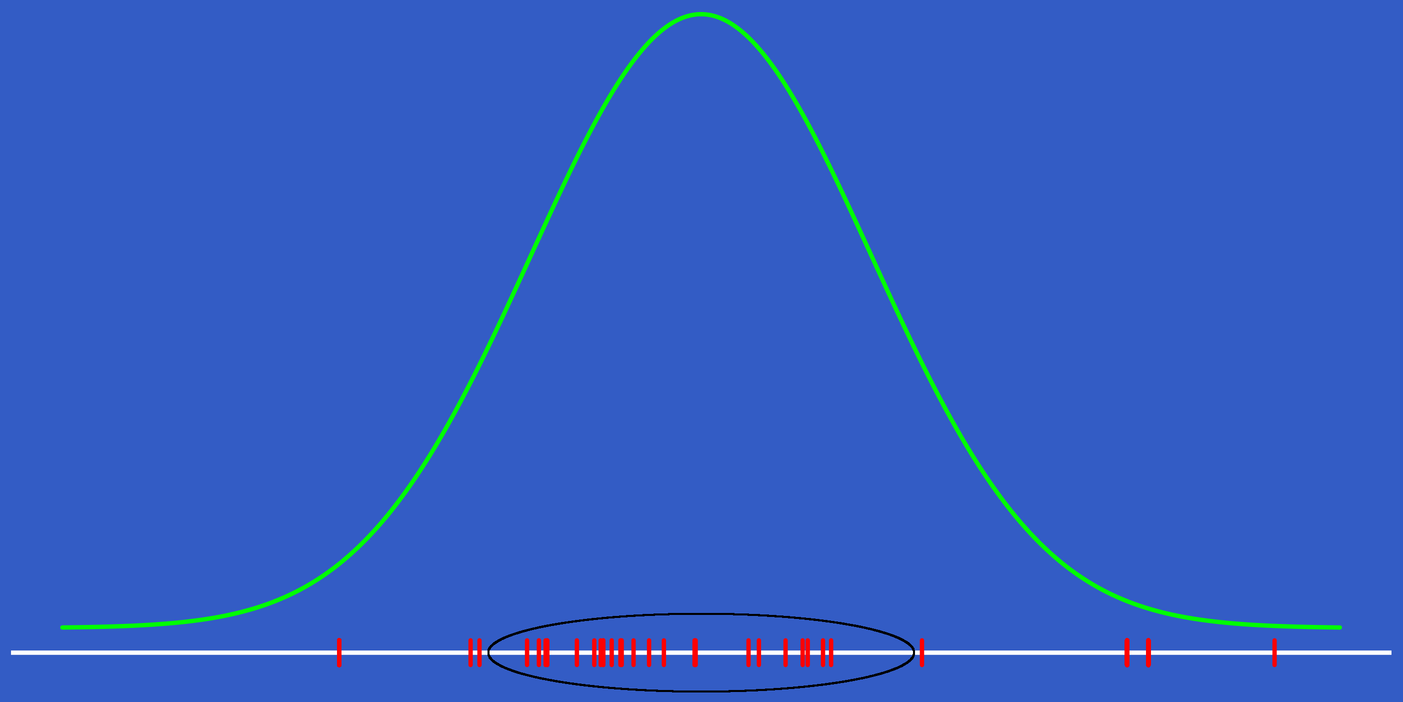
NCAR

Outline

- Part I - Two basic approaches to extreme value theory – block maxima, threshold models.
- Part II - Uncertainty, dependence, seasonality, trends.

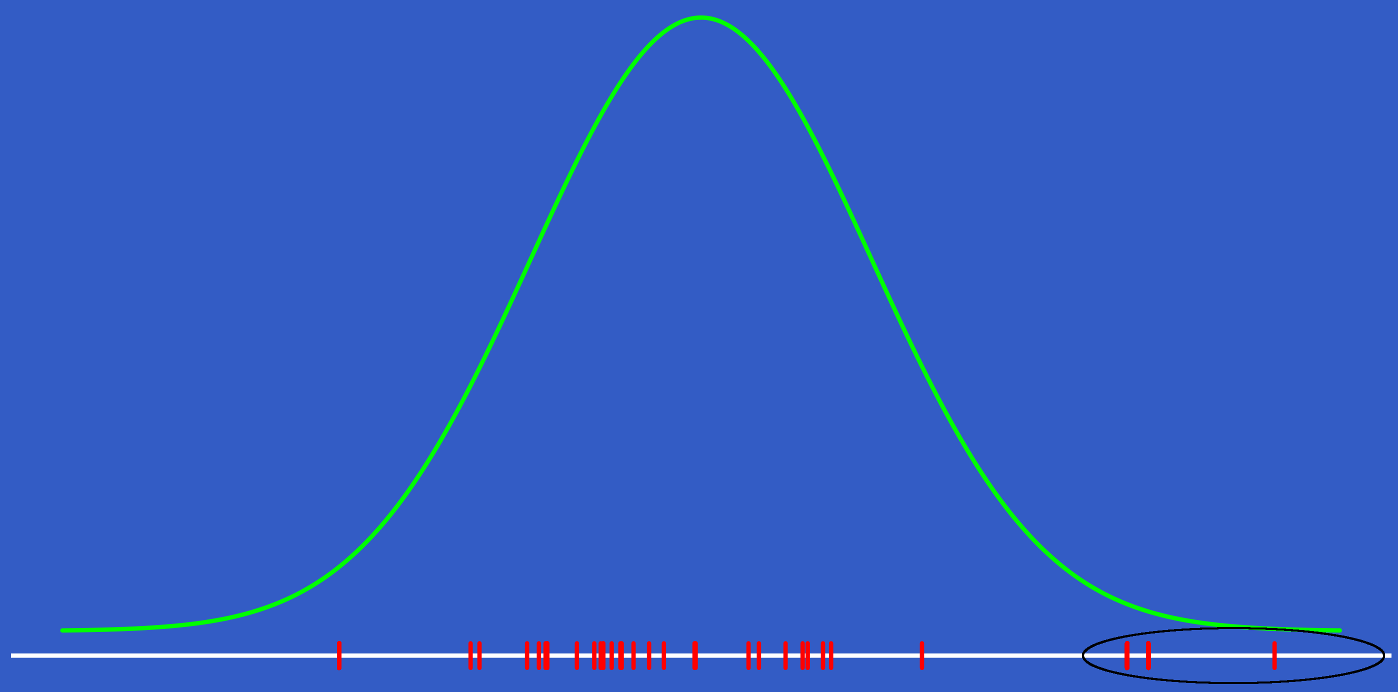
Fundamentals

- In classical statistics: model the AVERAGE behavior of a process.



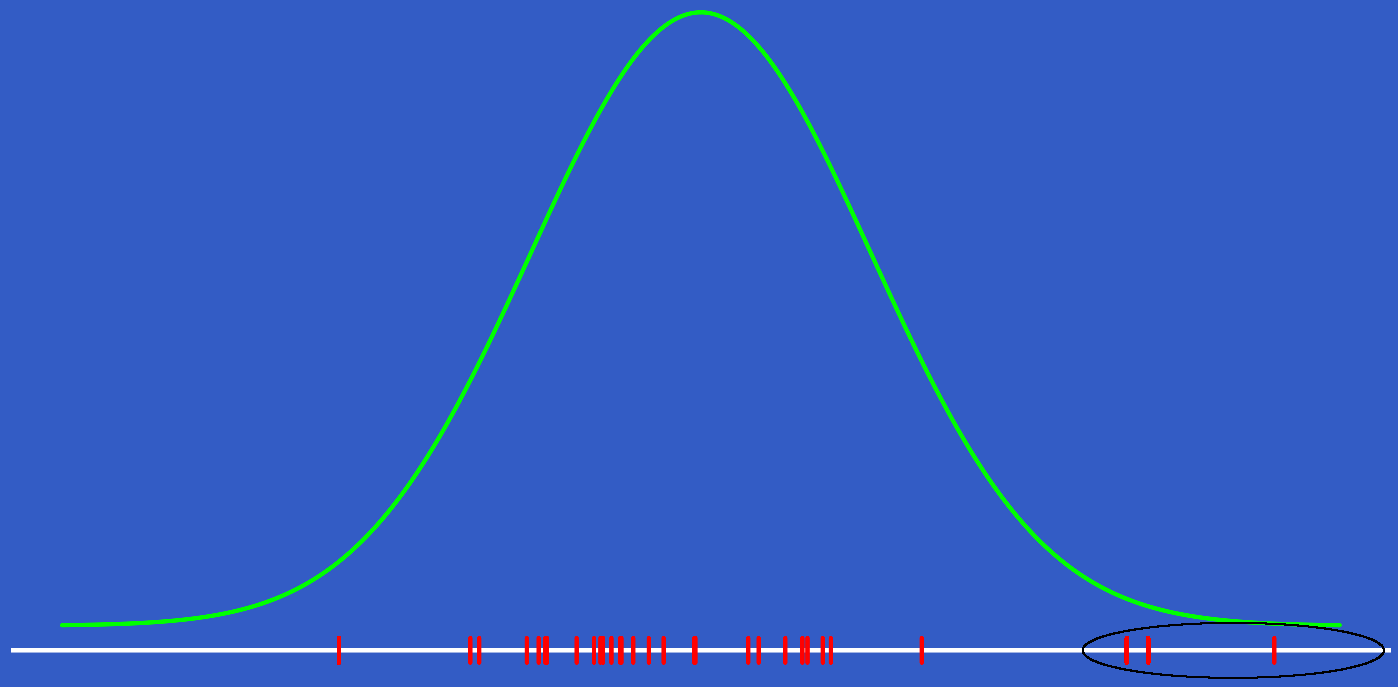
Fundamentals

- In extreme value theory: model the EXTREME behavior (the tail of a distribution).



Fundamentals

- In extreme value theory: model the EXTREME behavior (the tail of a distribution).



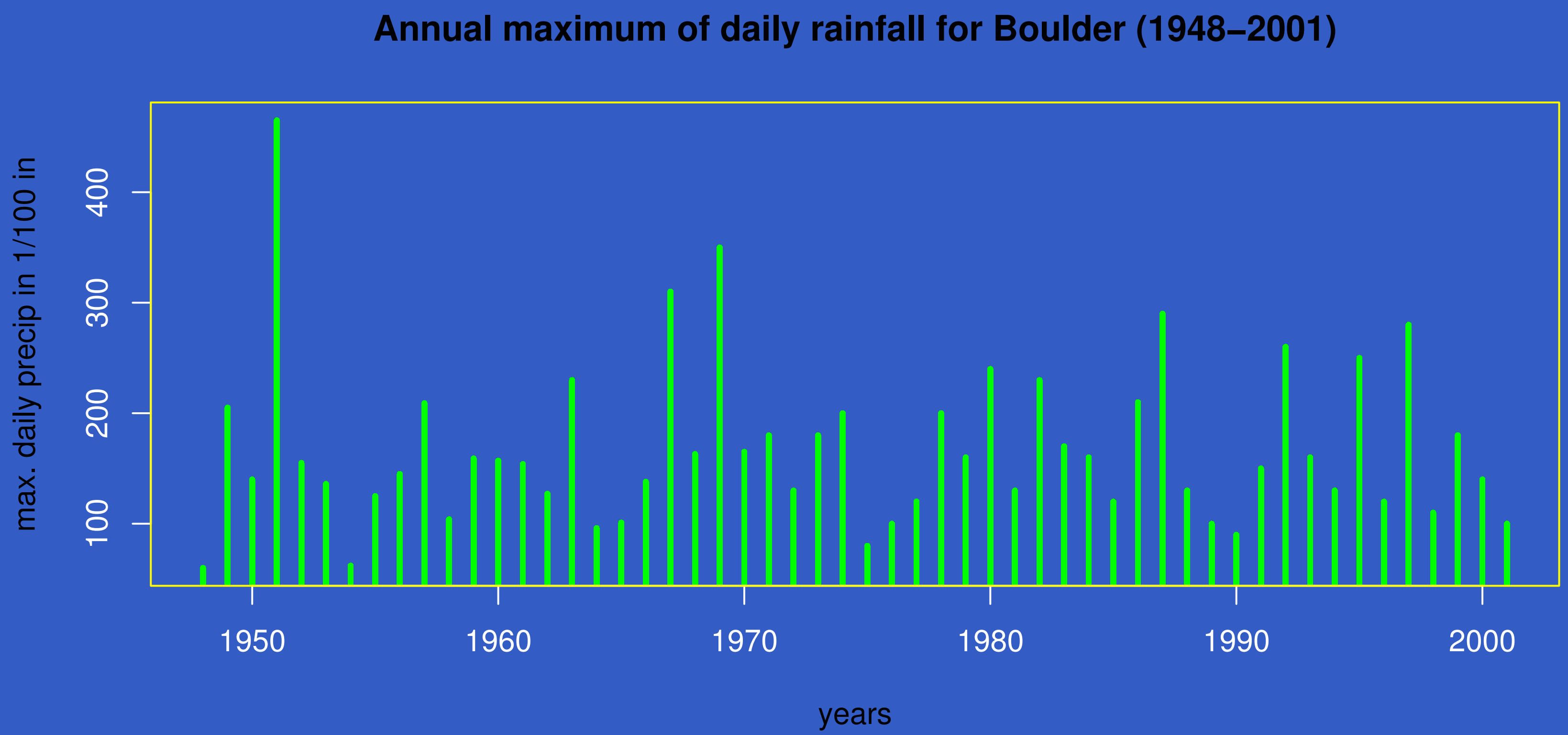
- Usually deal with very small data sets!

Different Approaches

- Block Maxima (GEV)
- Rth order statistic
- Threshold approach (GPD)
- Point processes

Block Maxima Approach

- Model extreme daily rainfall in Boulder
- Take "block maximum" – maximum daily precipitation for each year: $M_n = \max\{X_1, \ldots, X_{365}\}$
- 54 annual records (data points for M_n):



Block Maxima Approach

- The distribution of $M_n = \max\{X_1, \ldots, X_n\}$ converges to (as $n \to \infty$)

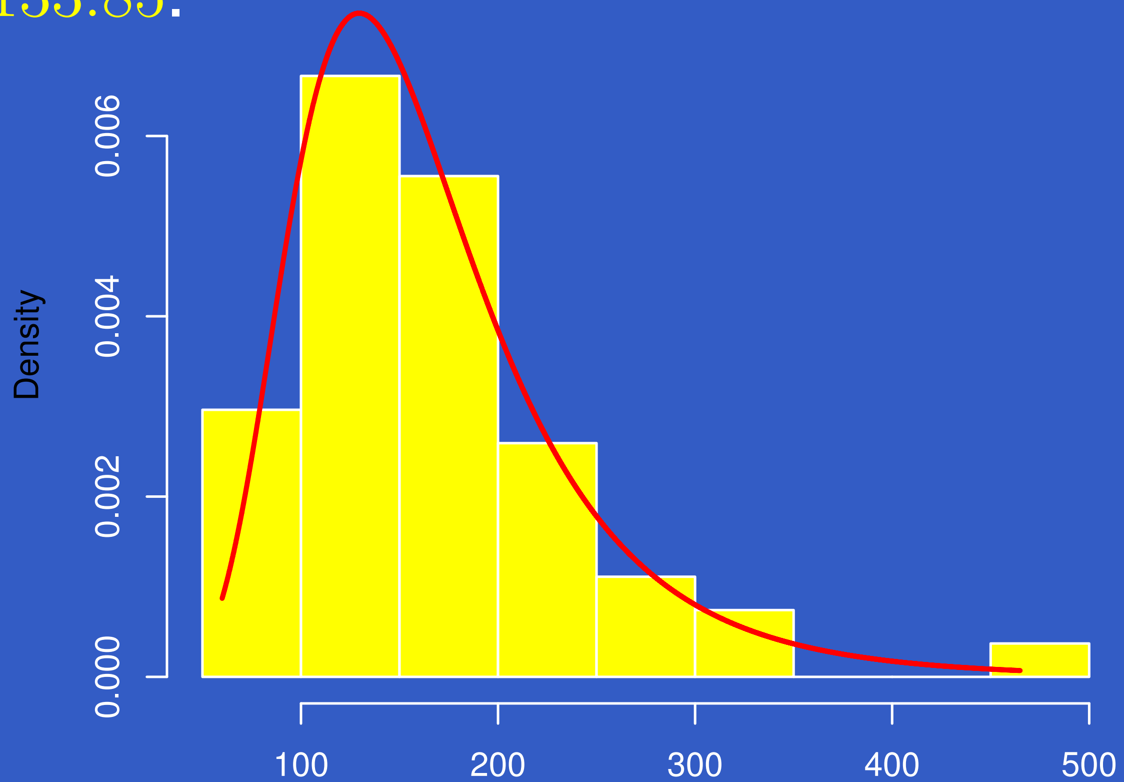
$$G(x) = \exp\{-[1 + \xi(\frac{x-\mu}{\sigma})]^{-\frac{1}{\xi}}\}.$$

$G(x)$ is called the "Generalized Extreme Value" (GEV) distribution and has 3 parameters:

  - shape parameter ξ
  - location parameter μ
  - scale parameter σ.

Fitting a GEV – Estimating Parameters

- Use the 54 annual records to fit the GEV distribution.
- Estimate the 3 parameters ξ, μ and σ with “maximum likelihood” (MLE) using statistical software (R).
- Get a GEV distribution with $\xi = 0.09$, $\mu = 50.16$, and $\sigma = 133.85$.



Fitting a GEV – Return Levels

- Often of interest: return level z_m

$$P(M > z_m) = \frac{1}{m}.$$

Expect every m^{th} observation to exceed the level z_m.
Or: at any point, there is a $1/m\%$ probability to exceed the level z_m.

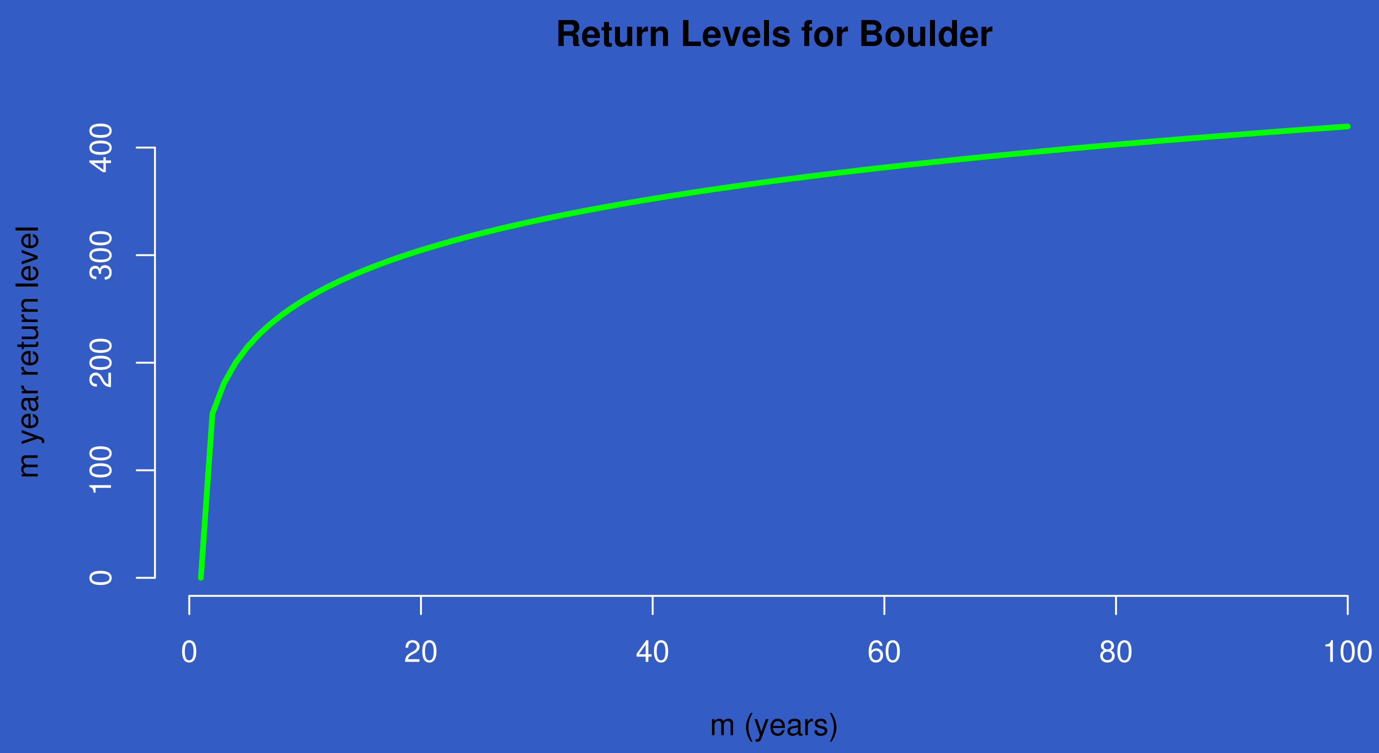
- Can be computed easily once the parameters are "known".
- E.g. $m = 100$, then $z_{100} = 420$, i.e. expect the annual daily maximum to exceed 4.2 inches every 100 years in Boulder.

Fitting a GEV – Return Levels

- Often of interest: return level z_m

$$P(M > z_m) = \frac{1}{m}.$$

Expect every m^{th} observation to exceed the level z_m.

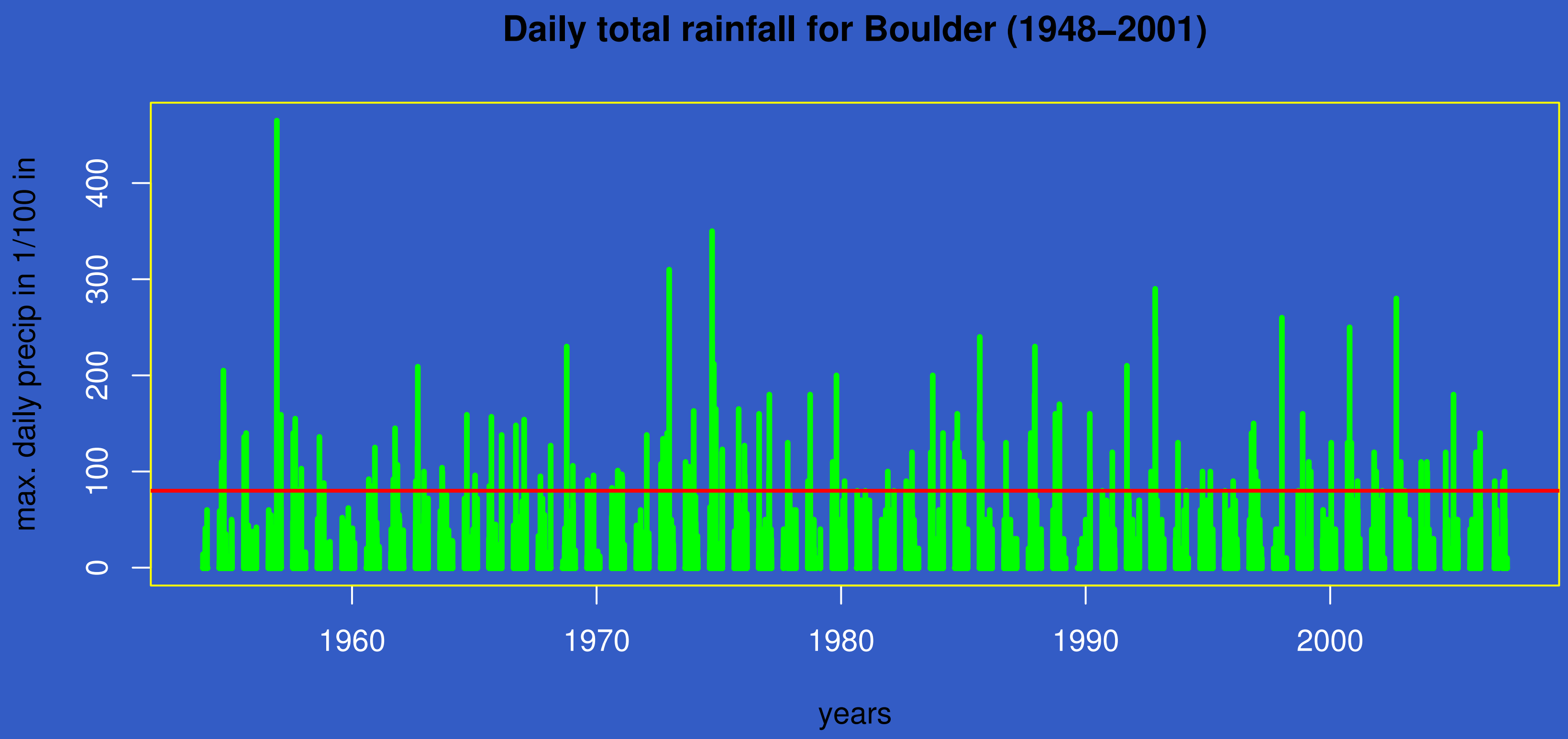


Fitting a GEV – Assumptions

- We did not need to know what the underlying distribution of each X_i, i.e. the daily total rainfall was.
- Underlying assumption: observations are "iid"
  - independently and
  - identically distributed.

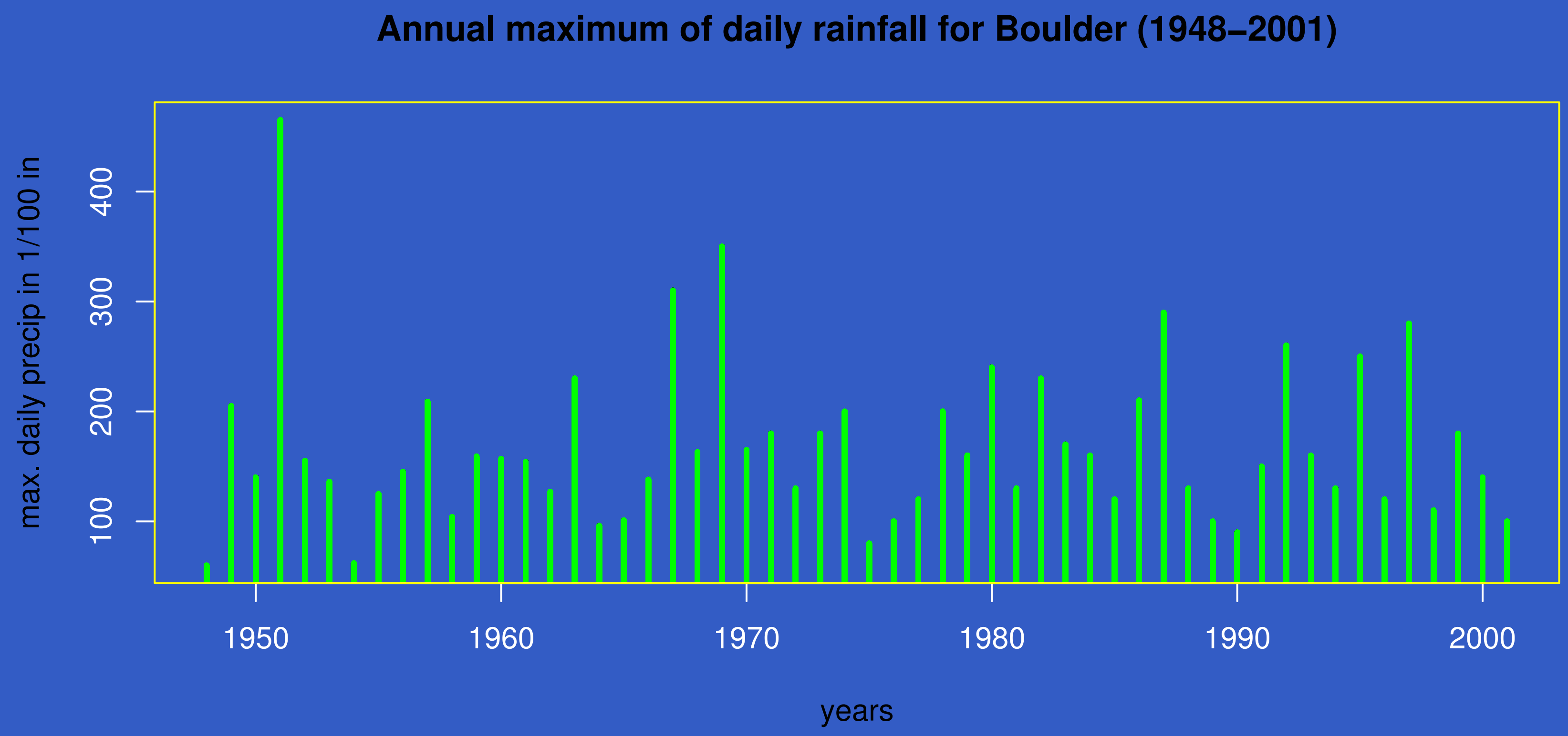
Threshold Models

- Model exceedances over a high threshold u – $X - u | X > u$.
- Daily total rainfall for Boulder exceeding 80 (1/100 in).
- Allows to make more efficient use of the data.



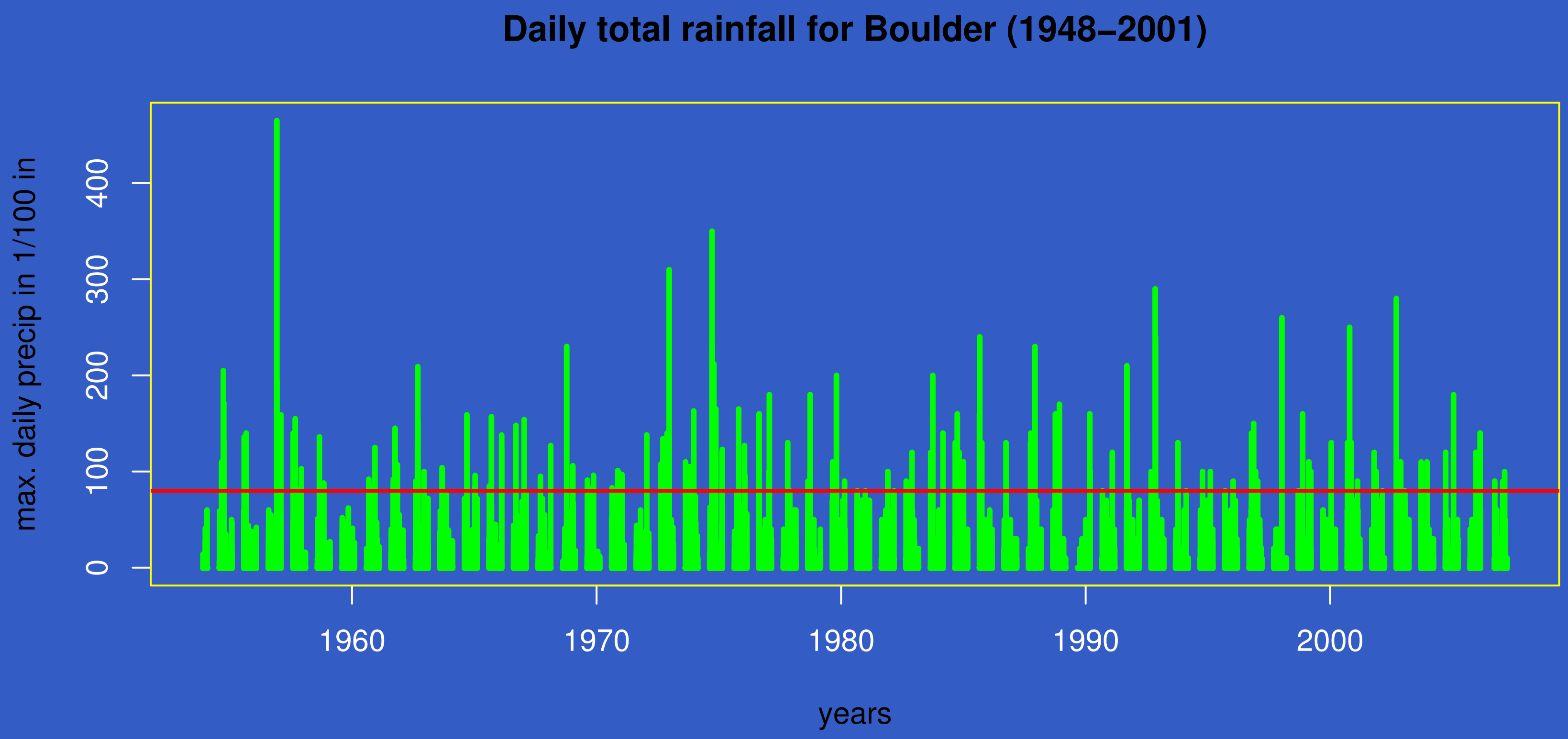
Threshold Models

- Model exceedances over a high threshold u – $X - u | X > u$.
- Daily total rainfall for Boulder exceeding 80 (1/100 in).
- Allows to make more efficient use of the data.



Threshold Models

- Model exceedances over a high threshold u – $X - u | X > u$.
- Daily total rainfall for Boulder exceeding 80 (1/100 in).
- Allows to make more efficient use of the data.



Threshold Models

- The distribution of $Y := X - u | X > u$ converges to (as $u \longrightarrow \infty$)

$$H(y) = 1 - (1 + \xi \frac{y}{\tilde{\sigma}})^{-\frac{1}{\xi}}.$$

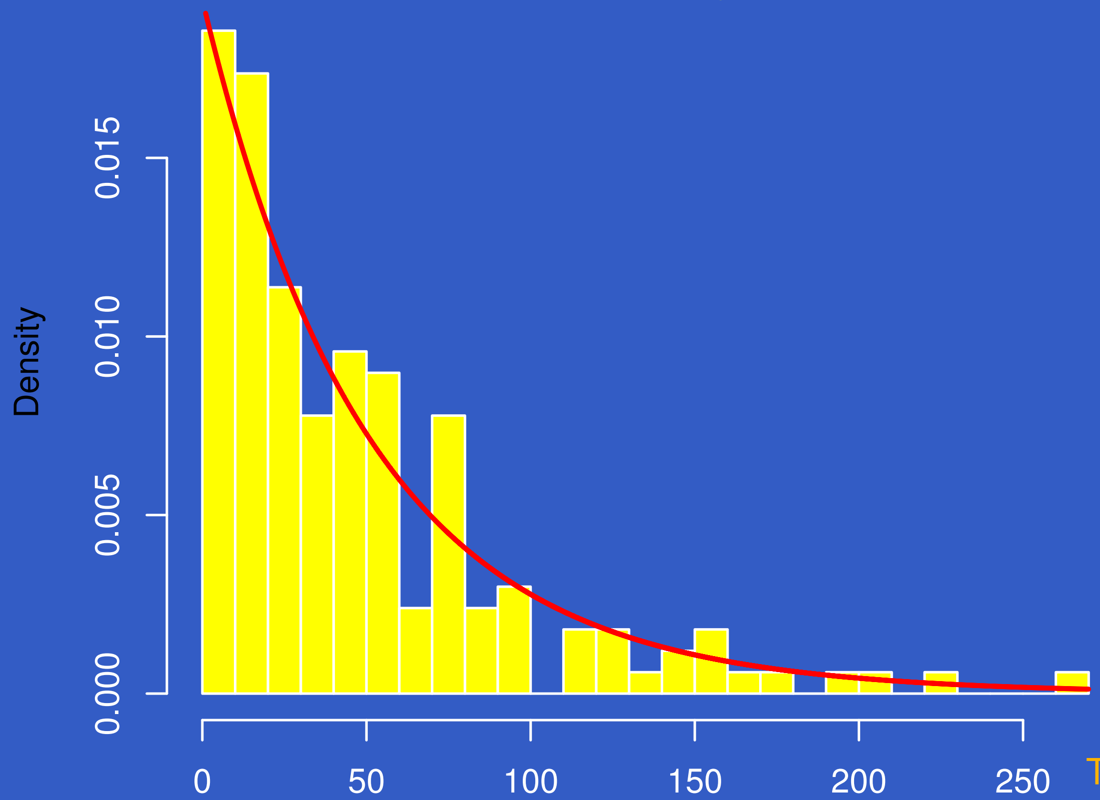
$H(y)$ is called the “Generalized Pareto” distribution (GPD) with 2 parameters.

  - shape parameter ξ
  - scale parameter $\tilde{\sigma}$.

The shape parameter ξ is the “same” parameter as in the GEV distribution.

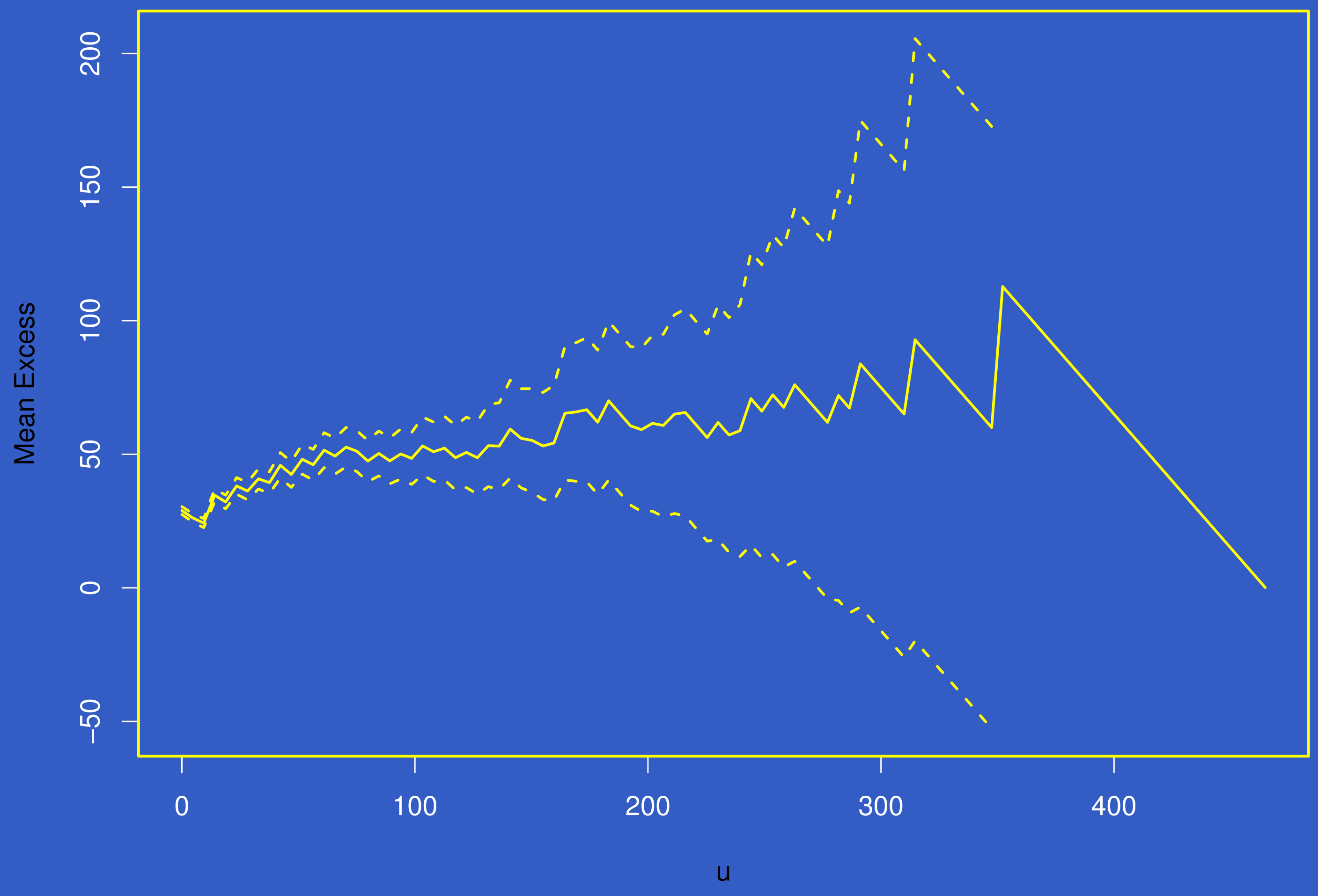
Fitting a GPD – Estimating Parameters

- Use the 184 exceedances over the threshold $u = 80$ to fit the GEV distribution.
- Estimate the 2 parameters ξ and σ (using "maximum likelihood" using statistical software (R).
- Get a GPD distribution with $\xi = 0.22$ and $\sigma = 51.46$.



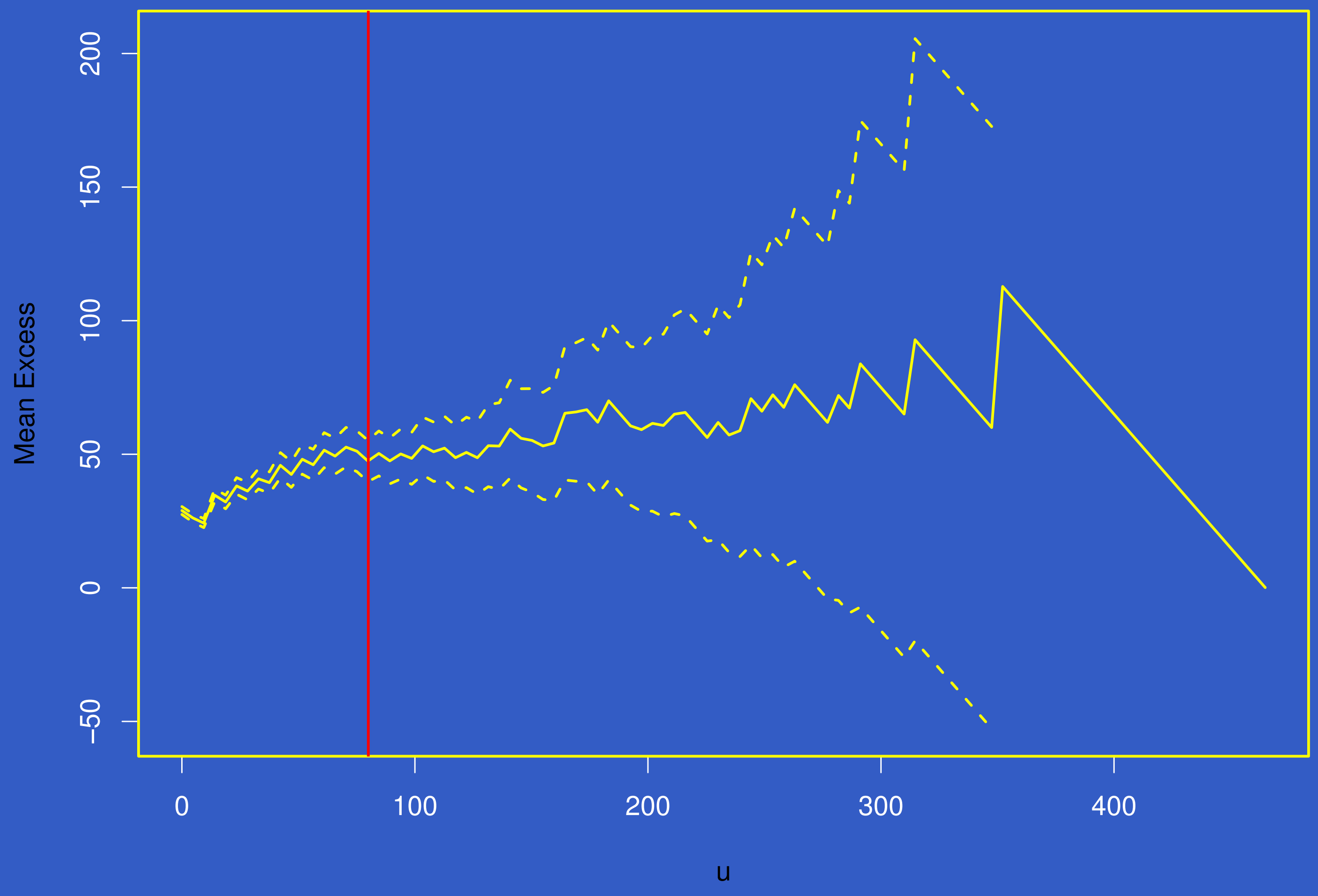
Fitting a GPD – Choosing a Threshold

Diagnostics: mean excess function – linear?



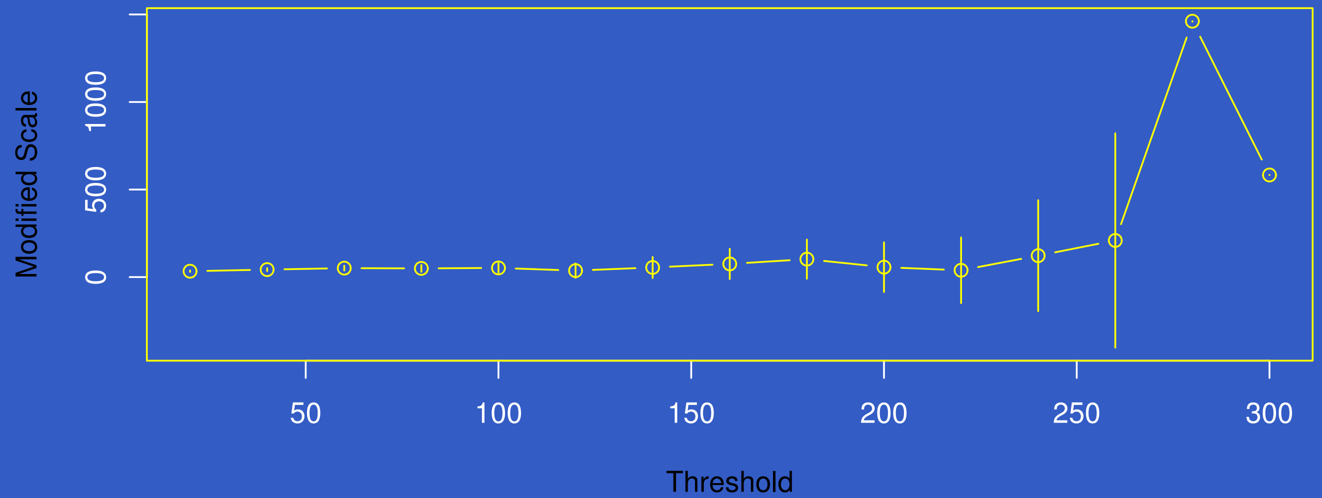
Fitting a GPD – Choosing a Threshold

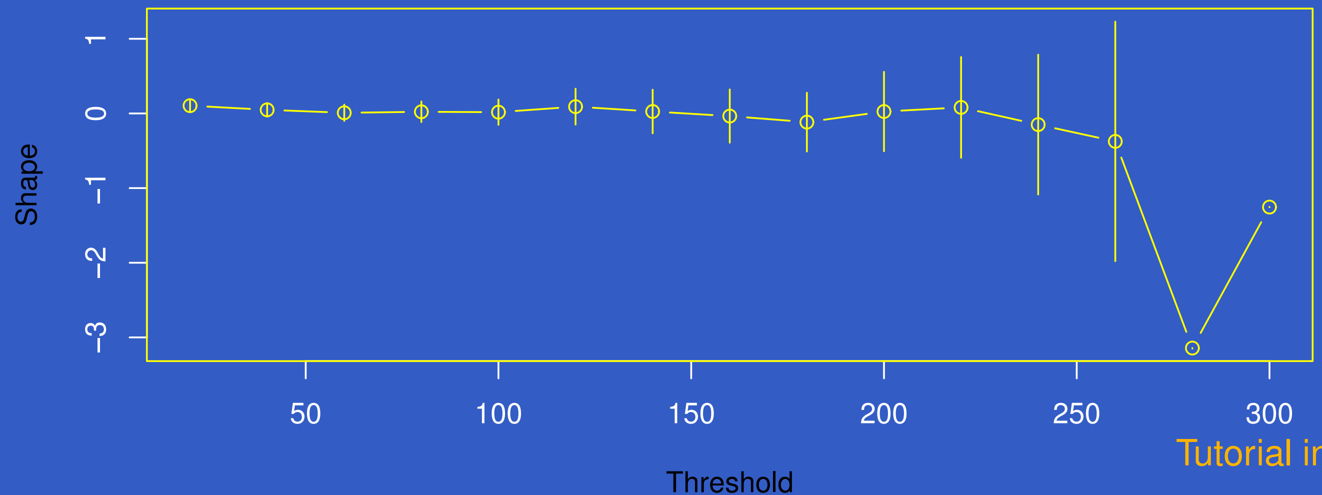
Diagnostics: mean excess function – linear?



Fitting a GPD – Choosing a Threshold

Diagnostics: shape and modified scale – constant?





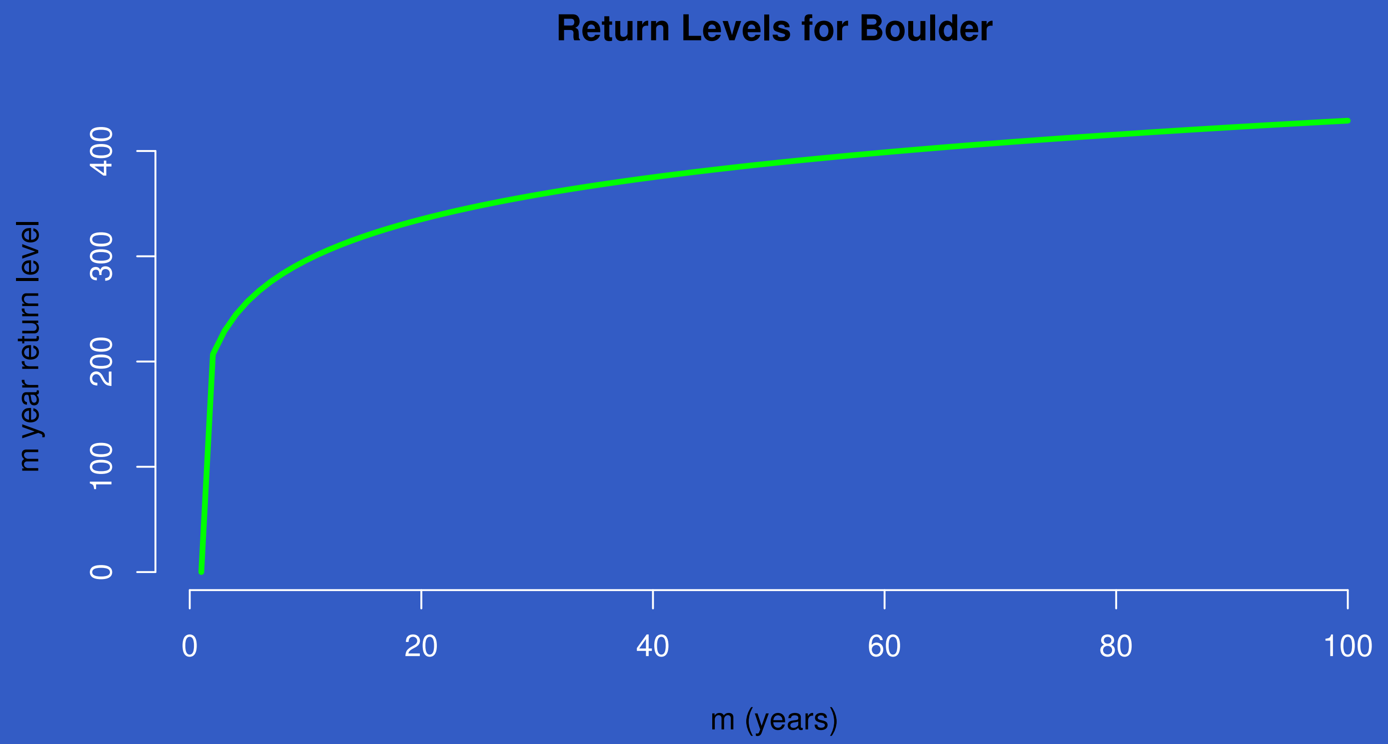
Fitting a GPD – Choosing a Threshold

Alternatively:

Choose the threshold u so that a certain percentage of the data lies above it (robust and automatic, but is the approximation valid?).

Fitting a GPD – Return Levels

- Compute 100-year return level for daily rainfall totals using the threshold approach: $z_{36500} = 429$, i.e. expect the daily total to exceed 4.29 inches every 100 years (36500 days).

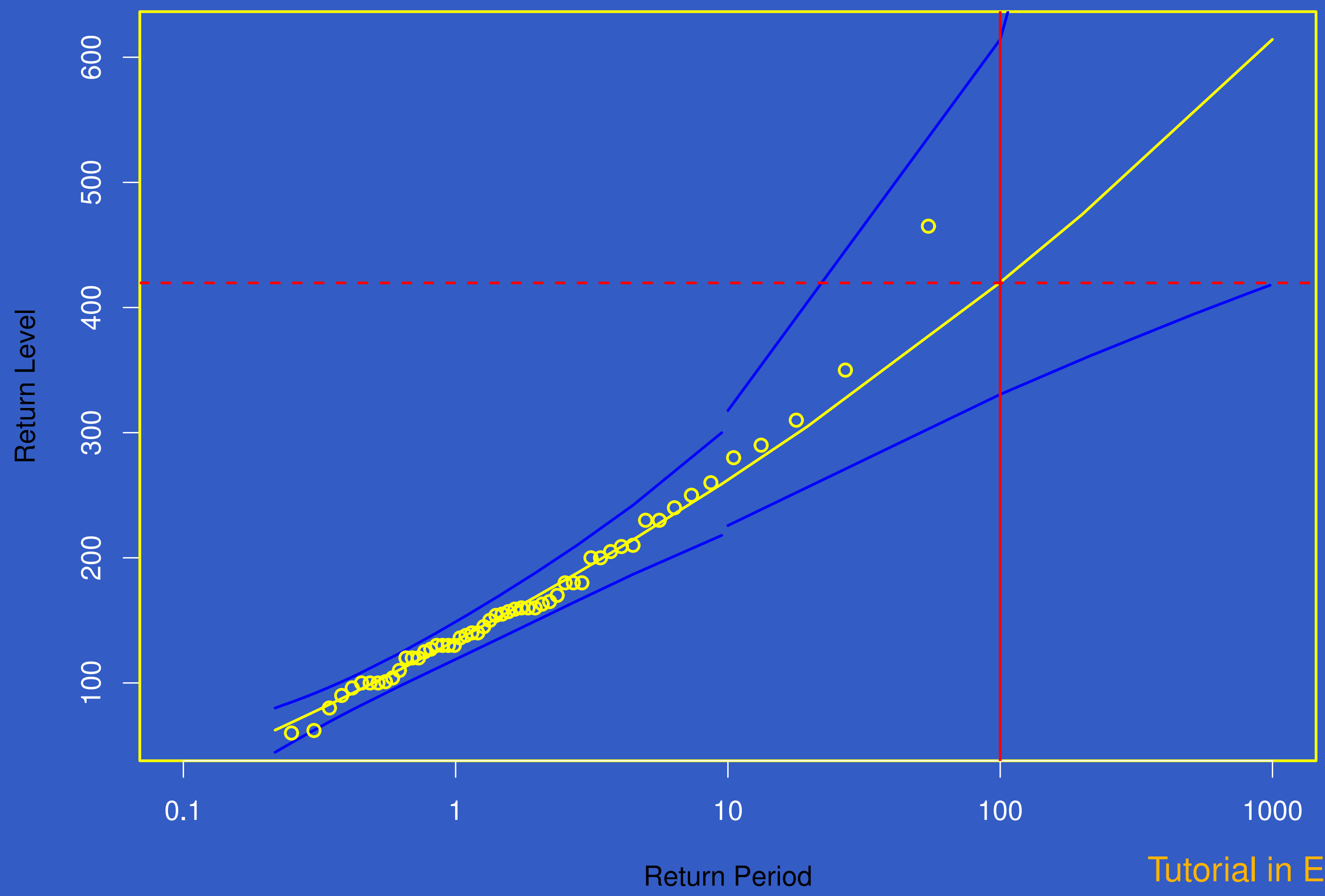


Uncertainty (GEV)

- Essentially, the maximum likelihood approach yields standard errors for the estimates and therefore confidence bounds on the parameters.
- From the GEV (block maxima) fit for the yearly maximum of daily precipitation for Boulder:
  - $\xi = 0.09$, 95% conf. interval is (-0.1,0.28).
  - $\sigma = 50.16$, 95% conf. interval is (38.77, 61.54).
  - $\mu = 133.85$, 95% conf. interval is (118.58,149.12).

Uncertainty (GEV)

- Essentially, the maximum likelihood approach yields standard errors for the estimates.
- These errors can be propagated to the return levels:



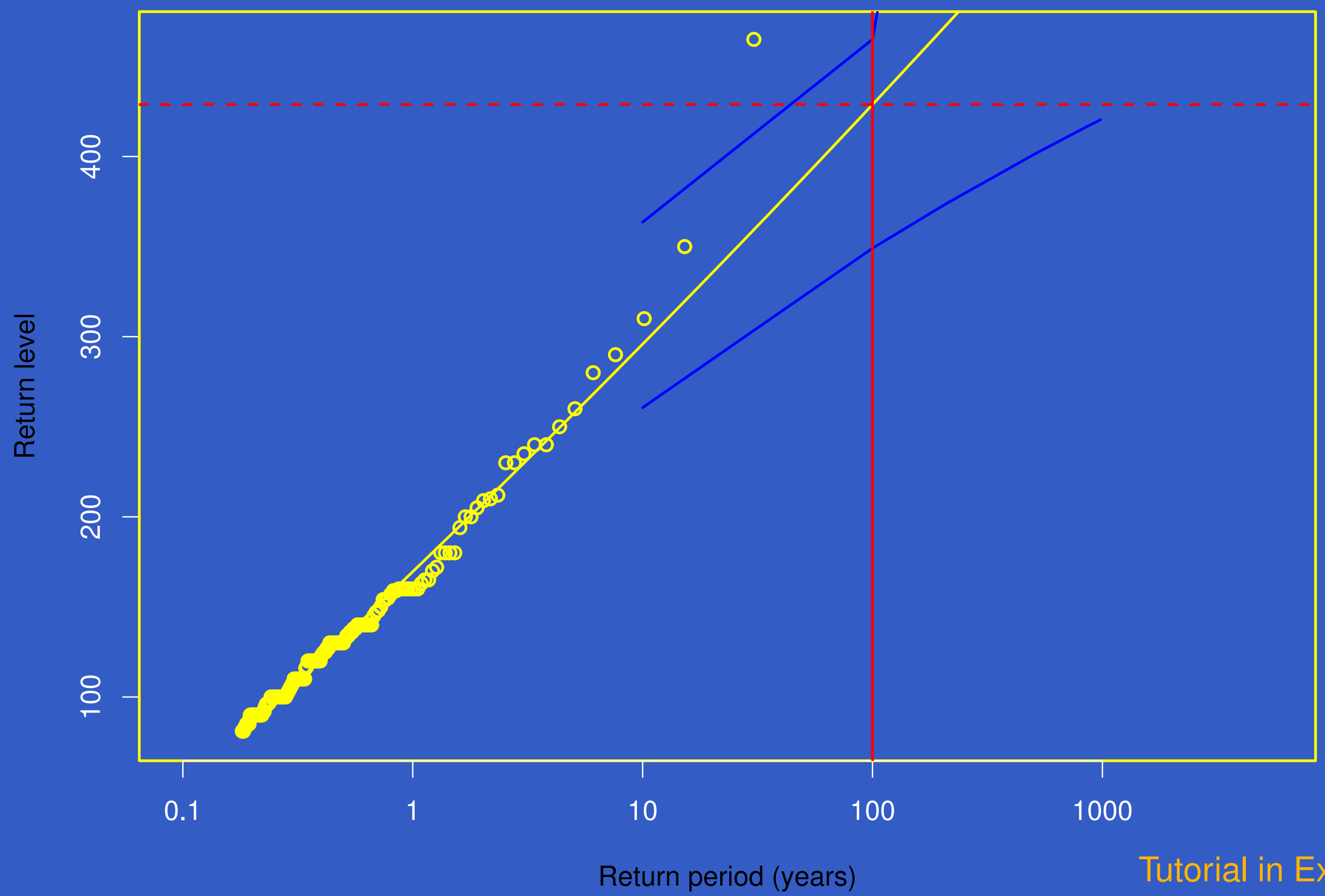
Uncertainty (GPD)

- More data means less uncertainty.
- From the GPD (threshold model) fit for daily precipitation in Boulder:

  - $\xi = 0.22$, 95% conf. interval is (-0.12,0.16).
  - $\sigma = 51.46$, 95% conf. interval is (40.70, 62.21).

Uncertainty (GPD)

- More data means less uncertainty.
- From the GPD (threshold model) fit for daily precipitation in Boulder:

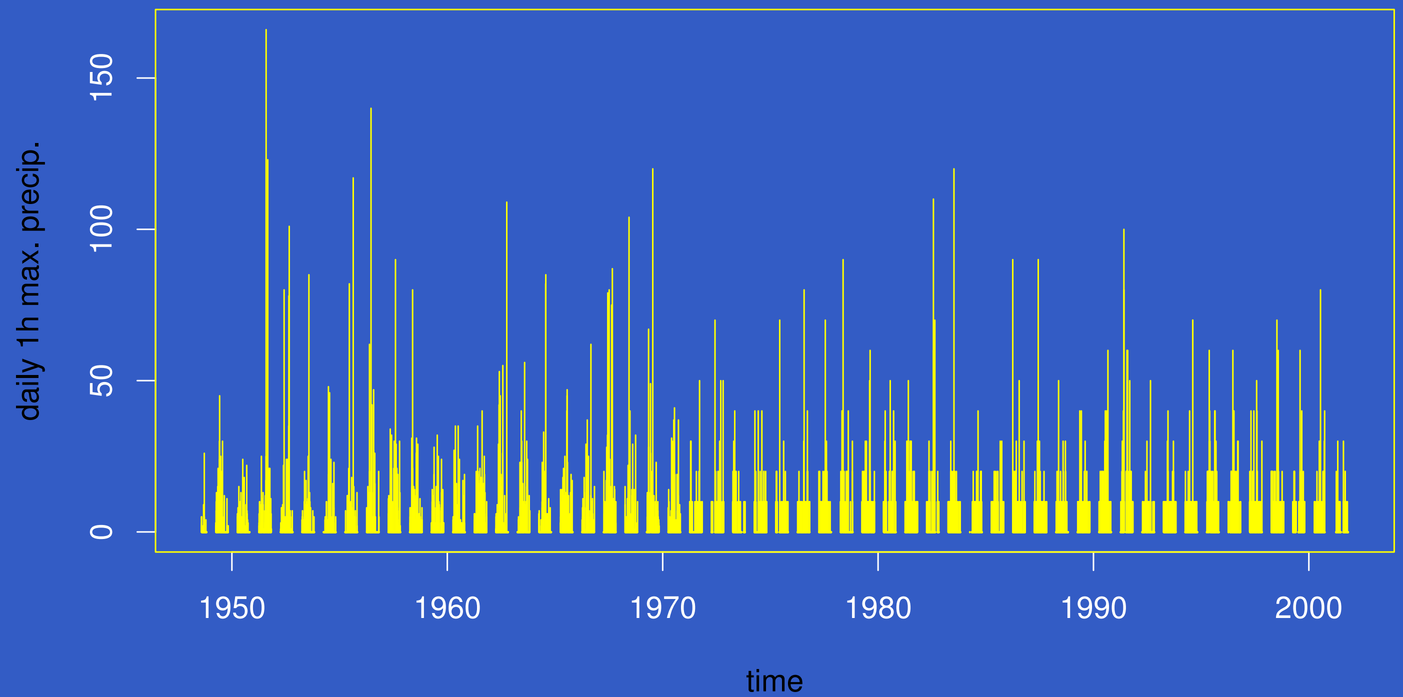


Dependence – Declustering

- For the GEV and GPD approximations to be valid, we assume independence of the data.
- If the data is dependent, can use declustering to “make” them independent.
- E.g. pick only one (the max) point in a cluster that exceeds a threshold.

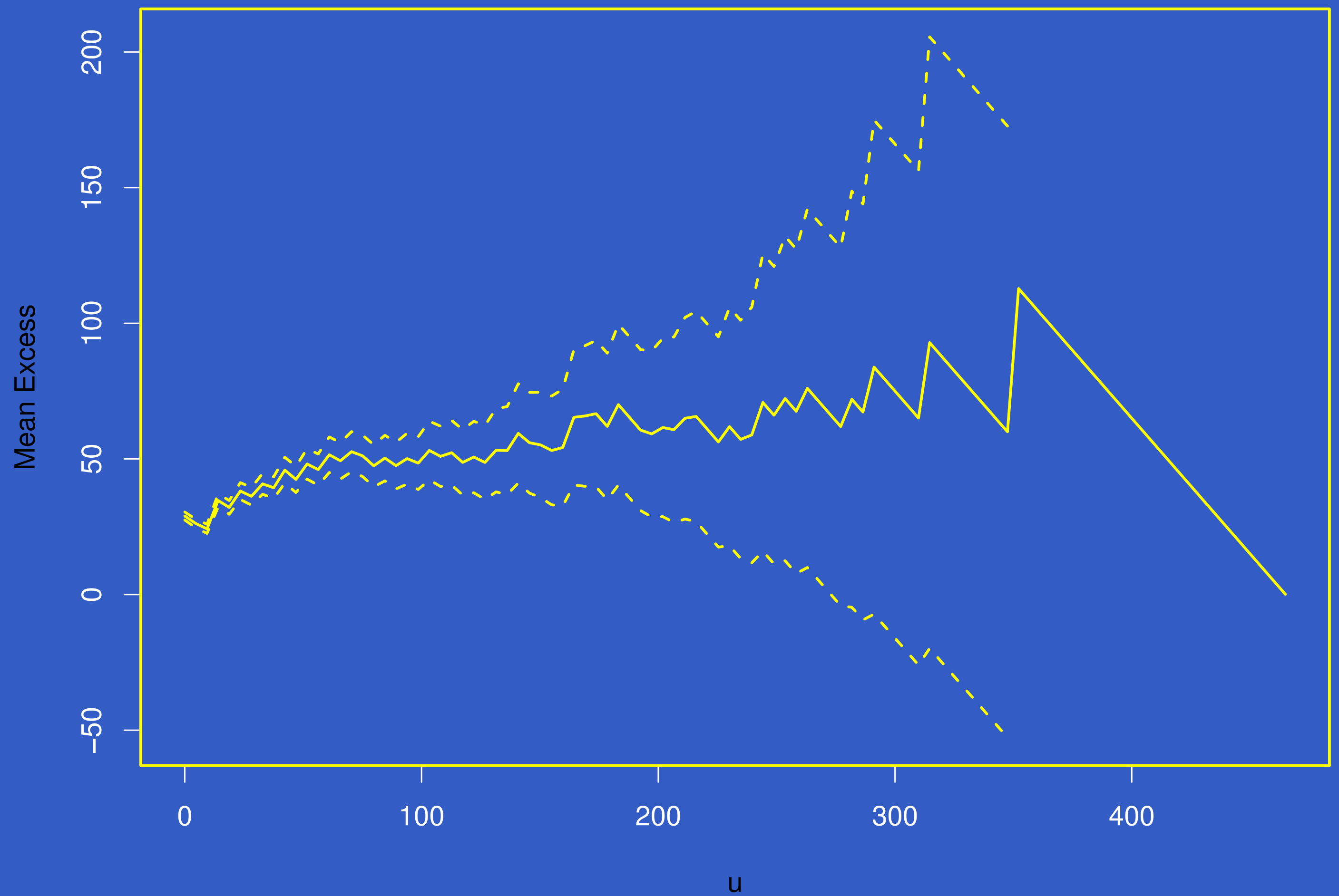
Dependence – Declustering

- Assume we want to make inference about hourly precipitation in Boulder.
- To decluster (instead of using 24 values for each day), we select only the maximum daily (1-h) record to fit the GPD model.



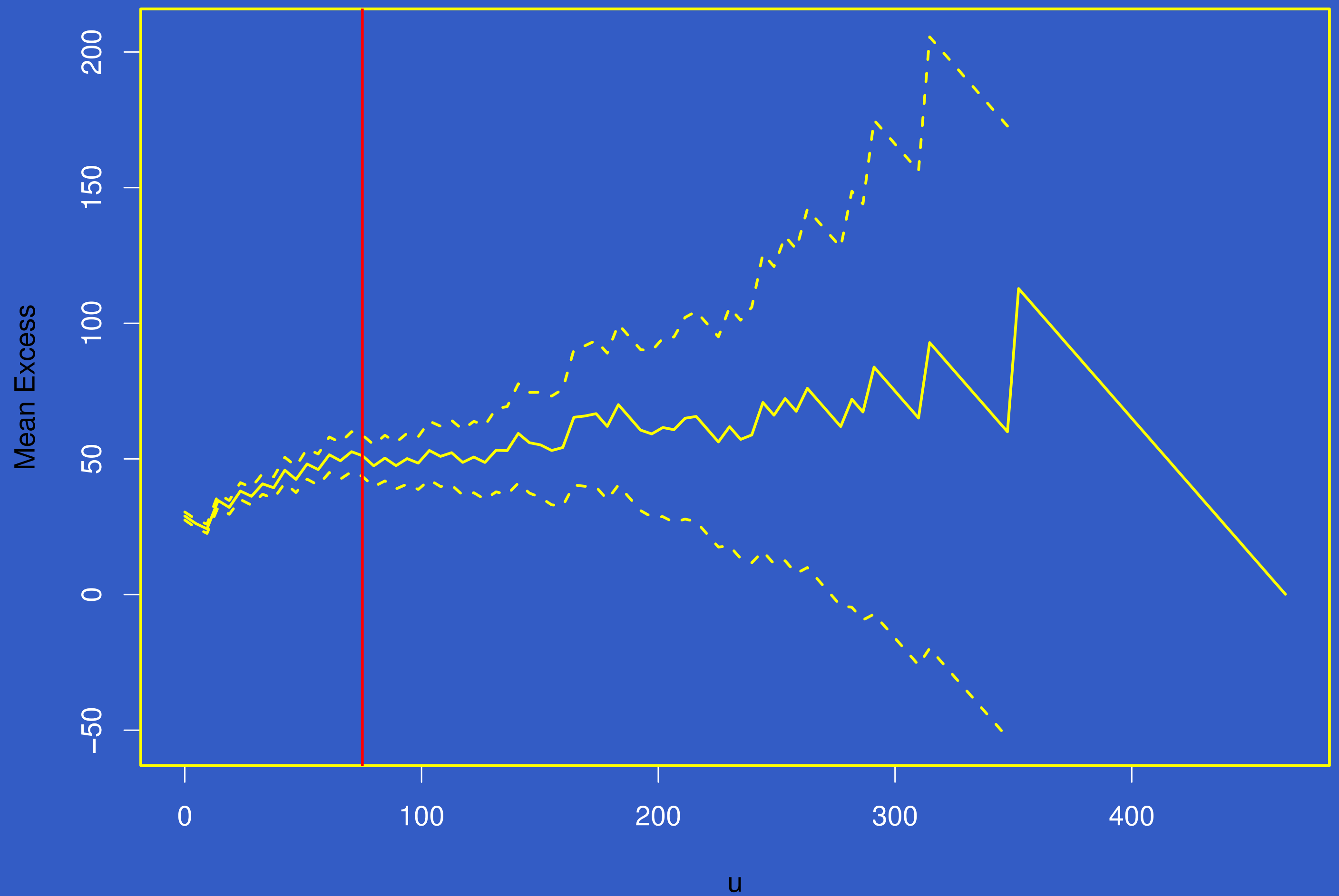
Dependence – (fitting the GPD)

Choosing a threshold – mean excess function as a diagnostic:

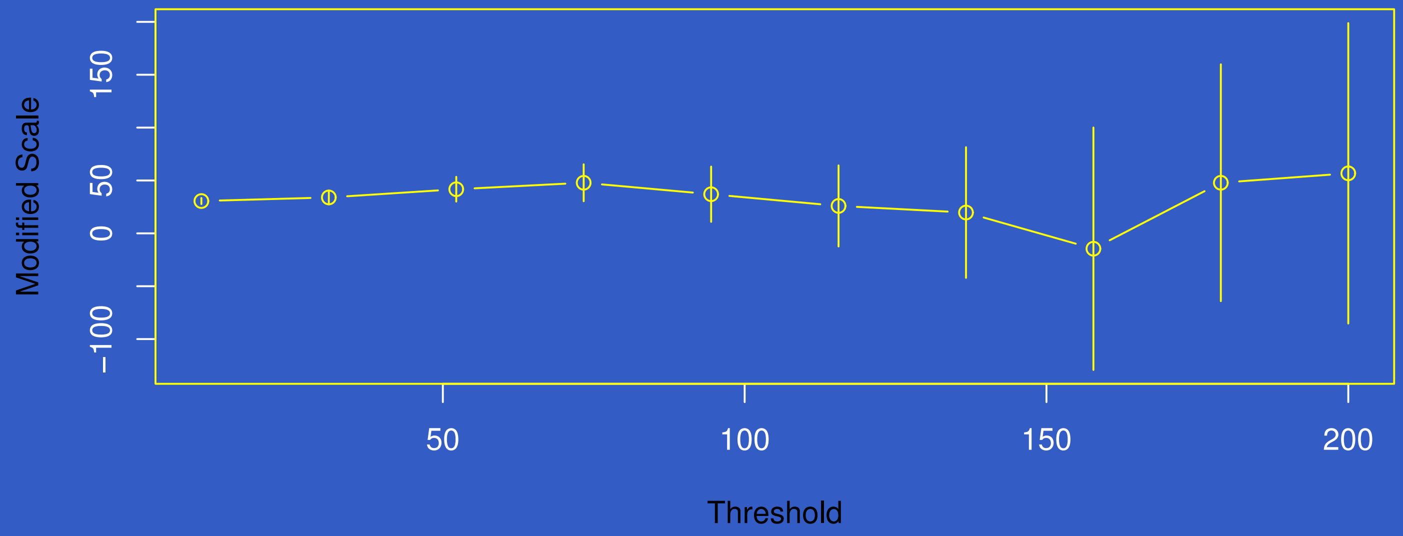


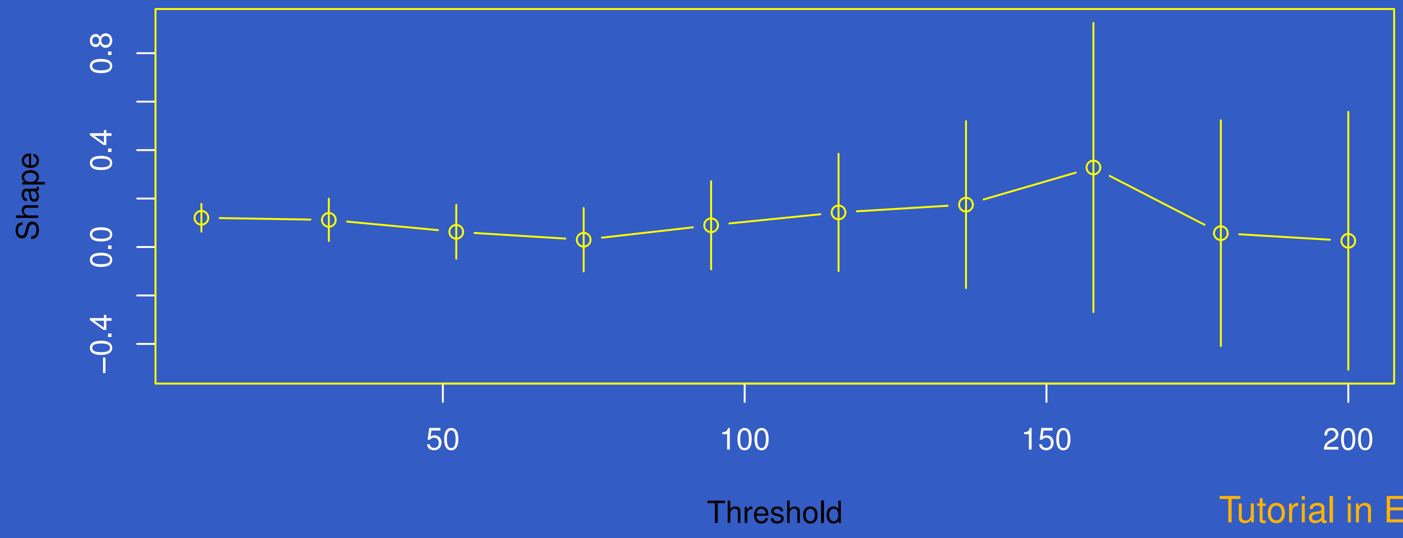
Dependence – (fitting the GPD)

Choosing a threshold – mean excess function as a diagnostic:



Dependence – (fitting the GPD)



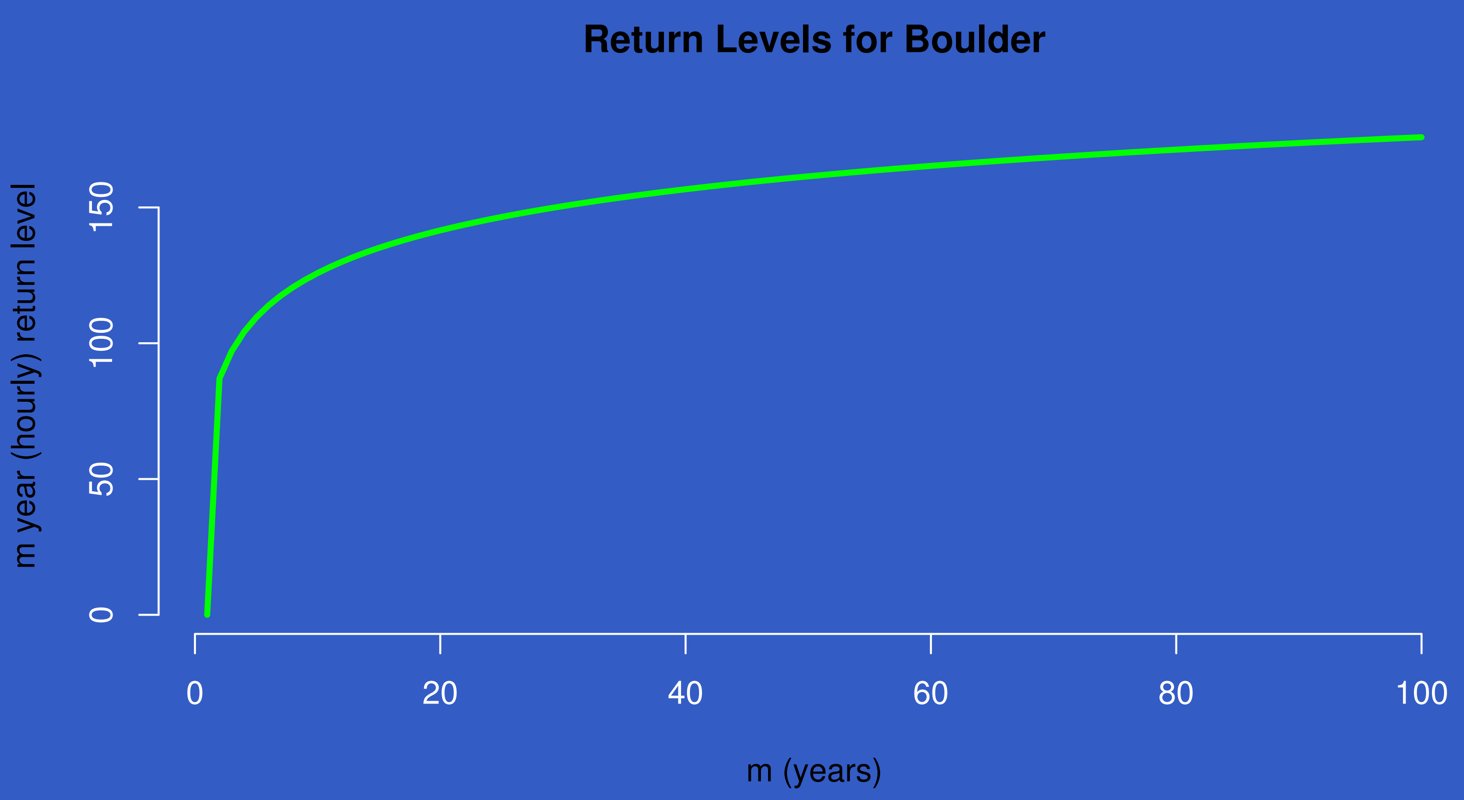


Dependence – (fitting the GPD)

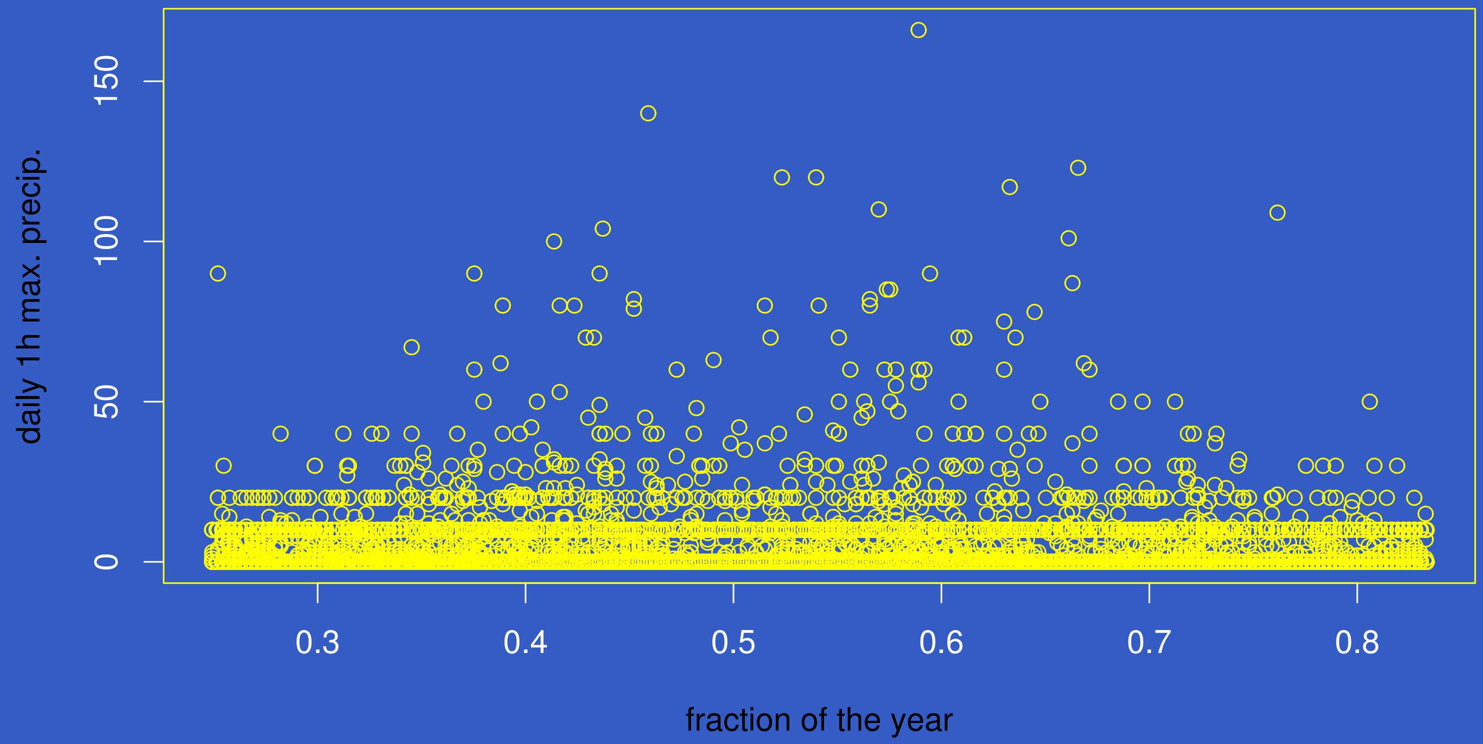
- $u = 75$ seems to be a good threshold using the diagnostics.
- But $u = 75$ only leaves 28 data points above the threshold.
- Use $u = 35$ instead (with 108 data points above the threshold) to get the following estimates:
  - $\xi = -0.05$, 95% conf. interval is (-0.27,0.15).
  - $\sigma = 27.98$, 95% conf. interval is (19.94, 36.02).
  - 100-year return level is $z_m = 185$, i.e. expect the hourly rainfall to exceed 1.85 inches every 100 years (10-year level is 1.36 inches.)

Dependence – (fitting the GPD)

- Use $u = 35$ (with 108 data points above the threshold) to fit a GPD model.

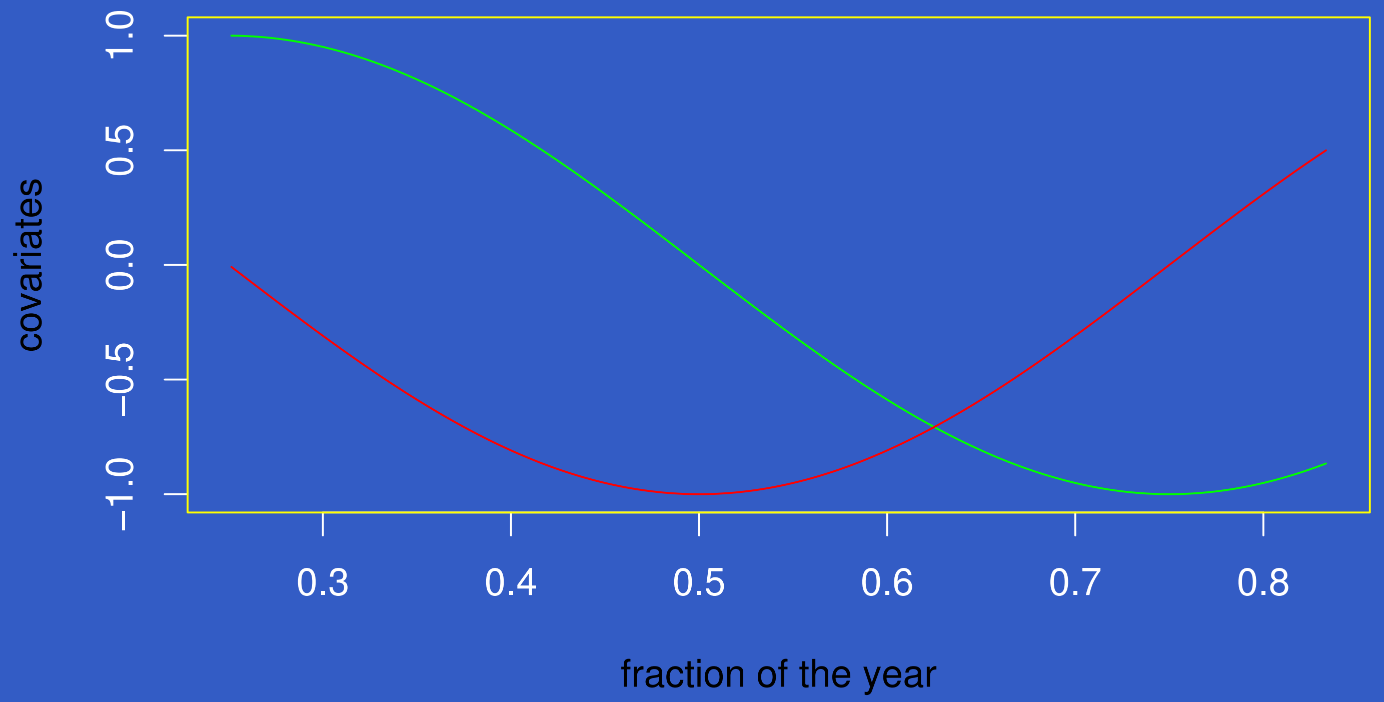


Seasonality



Seasonality

- To incorporate seasonality, link the scale parameter to covariates to describe the seasonal cycle.
- Use the covariates $X_1(t) = \sin(2\pi f(t))$ and $X_2(t) = \cos(2\pi f(t))$, where $f(t) =$fraction of the year for each day t.



Seasonality

- To incorporate seasonality, link the scale parameter to covariates to describe the seasonal cycle.
- Use the covariates $X_1(t) = \sin(2\pi f(t))$ and $X_2(t) = \cos(2\pi f(t))$, where $f(t) =$fraction of the year for each day t.
- Use an exponential link function to link the covariates to the scale parameter:

$$\tilde{\sigma}(t) = \exp(\beta_0 + \beta_1 X_1(t) + \beta_2 X_2(t)).$$

Fit a GPD with density

$$GPD\left(\xi, \tilde{\sigma}(t) = \exp(\beta_0 + \beta_1 X_1(t) + \beta_2 X_2(t))\right).$$